DOCUMENT RESUME

ED 390 925 TM 024 372

AUTHOR	Zwick, Rebecca; Thayer, Dorothy T.
TITLE	A Comparison of the Performance of Graduate and Undergraduate School Applicants on the Test of Written English. TOEFL Research Reports Report 50.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-15
PUB DATE	May 95
NOTE	41p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Biological Sciences; Comparative Analysis; *English (Second Language); Higher Education; High Schools; *High School Students; Humanities; Language Proficiency; *Language Tests; Majors (Students); Outcomes of Education; Physical Sciences; *Scores; Student Characteristics; Test Results; *Undergraduate Students
IDENTIFIERS	Test of English as a Foreign Language; *Test of Written English

ABSTRACT

The performance of graduate and undergraduate school applicants on the Test of Written English (TWE) was compared for each of 66 data sets, dating from 1988 to 1993. The analyses compared the average TWE score for graduates and undergraduates after matching examinees on the total score on the Test of English as a Foreign Language (TOEFL). The main finding was that, for matched examinees, undergraduate TWE means were higher than graduate means in 63 of the 66 data sets. Although these standardized mean differences (SMDs) never exceeded 0.3 of a TWE score point, the results are noteworthy because they give a different picture than do simple comparisons of means for unmatched graduates and undergraduates, which show higher mean TWE scores for graduates in the majority of cases. Of the 9 SMDs exceeding 0.2, 8 were in Region 1 (Asia and the Pacific) between October 1988 and May 1992. Effects of the examinees' intended fields of graduate study were also investigated. Applicants to programs in the physical and biological sciences tended to have lower TWE scores than those in the social sciences, and graduates in the physical sciences tended to have lower scores than those in the humanities. Reasons for these differences are discussed. (Contains 11 tables and 32 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-95-15

ED 390 925

TOEFL

TEST OF ENGLISH AS A FOREIGN LANGUAGE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Research Reports

REPORT 50
MAY 1995

A Comparison of the Performance of Graduate and Undergraduate School Applicants on the Test of Written English

Rebecca Zwick
Dorothy T. Thayer

Educational Testing Service

TM024512

2

BEST COPY AVAILABLE

A Comparison of the Performance of Graduate and Undergraduate School Applicants on the Test of Written English

Rebecca Zwick
Dorothy T. Thayer

Educational Testing Service
Princeton, New Jersey

RR-95-15

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Copyright © 1995 by Educational Testing Service. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRE, TOEFL, the TOEFL logo, TSE, and TWE are registered trademarks of Educational Testing Service.

Advanced Placement Program and AP are registered trademarks of the College Entrance Examination Board.

Abstract

The performance of graduate and undergraduate school applicants on the Test of Written English (TWE®) was compared for each of 66 data sets, dating from 1988 to 1993. The analyses compared the average TWE score for graduates and undergraduates after matching examinees on the TOEFL® total score. The main finding was that, for matched examinees, undergraduate TWE means were higher than graduate means in 63 of the 66 data sets. Although these standardized mean differences (SMDs) never exceeded 0.3 of a TWE score point (with standard errors that were typically between 0.01 and 0.02), the results are noteworthy because they give a different picture than do simple comparisons of means for unmatched graduates and undergraduates, which show higher mean TWE scores for *graduate* applicants in the majority of cases. Of the nine SMDs exceeding 0.2, eight were in Region 1 (Asia and the Pacific) between October 1988 and May 1992. In evaluating these findings, the effects of the examinees' intended field of graduate study were investigated. For groups matched on TOEFL score, applicants to graduate programs in the physical and biological sciences tended to have lower TWE means than undergraduates and graduates in the social sciences. Graduate students in the physical sciences tended to have lower TWE means than matched graduates in the humanities. Graduate-undergraduate differences in TWE performance may result in part from a greater concentration of graduate applicants in scientific fields. Another hypothesis is that undergraduates are more likely to have recently participated in intense English writing instruction. Region 1 may show larger graduate-undergraduate differences because of greater demographic disparities in that region between these two groups of examinees.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1994-95) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
James Dean Brown	University of Hawaii
Carol Chapelle	Iowa State University
Joan Jamieson	Northern Arizona University
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Acknowledgments

We would like to thank Gwyneth Boodoo, Brent Bridgeman, Philip Everson, Gordon Hale, Liz Hamp-Lyons, Robert Kantor, Diana Marr, and Jacqueline Ross for providing comments on our findings; the staff from Testing Programs Systems and SHEP Statistical Analysis, especially Patricia Carey and Diana Marr, for helping us to assemble the needed data and answering our many questions; Jo-lin Liang, for creating tables and reviewing the report; and Martha Byelich and Tonia Williams for assisting with manuscript preparation.

Table of Contents

Page

Overview 1

Previous Research on Group Differences in Essay Performance 3

Methods 3
- Mantel's Statistical Procedure for Ordered Variables 3
- Descriptive Index of Differential Performance 4
- Hypothetical Example 5

TWE Analyses 6
- TWE Data Sets 6
  - Editing 7
- Preliminary Analyses 7
  - TWE Scores 8
  - TOEFL Scores 8
- Group Comparisons 9
  - Choice of TOEFL Total Score as a Matching Variable 9
  - Graduate-undergraduate Comparisons 10
  - Field-of-study Comparisons 11
  - Supplementary Analyses of the Effects of Gender, Native Country, and Native Language 12

Summary and Discussion 14

References 17

Tables 21

Tables

Page

Table 1 Hypothetical Example of the Mantel Approach and SMD Statistic: Frequencies of Graduates and Undergraduates Receiving Each TWE Score 21

Table 2 Examinee Data Included in the Study 22

Table 3 Preliminary Analysis - Region 1 23

Table 4 Preliminary Analysis - Region 2 24

Table 5 Preliminary Analysis - Region 3 25

Table 6 Results of Graduate - Undergraduate Comparisons - Region 1 26

Table 7 Results of Graduate - Undergraduate Comparisons - Region 2 27

Table 8 Results of Graduate - Undergraduate Comparisons - Region 3 28

Table 9 Summary of All Group Comparisons - Region 1 29

Table 10 Summary of All Group Comparisons - Region 2 30

Table 11 Summary of All Group Comparisons - Region 3 31

Overview

The Test of Written English (TWE®), introduced in 1986, is the essay component of the Test of English as a Foreign Language (TOEFL®), a multiple-choice measure of English language proficiency for nonnative speakers of English. TWE is intended to "provide information about an examinee's ability to generate and organize ideas on paper, to support those ideas with evidence or examples, and to use the conventions of standard written English" (*TOEFL Test of Written English Guide*, 1992, p. 5). Currently, TWE is offered five times a year.

The issue of performance differences between graduate and undergraduate school applicants across TWE essay prompts and administrations was recently identified as a high-priority research topic by TWE Committee members, who believed that a detailed analysis of graduate and undergraduate TWE scores could help to determine whether different forms of the test are needed for graduate and undergraduate examinees. As the *TOEFL Test of Written English Guide* (1992, p. 17) indicates, the overall mean TWE score is nearly identical for graduate applicants (3.75) and undergraduate applicants (3.76). To further investigate the performance of these two groups, we conducted a more focused analysis of TWE responses. Our analyses followed three principles, intended to maximize the interpretability of the results.

First, comparisons were made separately for each essay prompt. As noted by Golub-Smith, Reese, and Steinhaus (1993), "different formats, topics, and topic types might elicit different writing performances from the same examinee or may promote successful performance for one examinee while impeding successful performance for another" (p. 5). In general, findings about the effects of the topic and structure of prompts on essay scores have been inconsistent. Recent reviews of this research appear in Brown, Hilgers, and Marsella (1991), Hamp-Lyons (1990), and Huot (1990). On TWE, Golub-Smith et al. found that administering different prompts to equivalent samples of examinees yielded some differences in TWE score distributions that were considered large enough to be of practical importance.

Second, comparisons were made for graduate and undergraduate examinees who were matched in terms of a separate measure of English language proficiency. Simply comparing the mean TWE scores for graduates and undergraduates could be misleading in the sense that the two groups may differ in overall English language proficiency. For example, if the graduate TWE mean exceeds the undergraduate mean, this may result from a general superiority in English proficiency for the graduates, rather than from any superior skill in responding to the particular TWE

prompt. Comparing TWE results for graduates and undergraduates who are matched in English proficiency allows prompt-specific group differences to be revealed. (For reasons described in a later section, TOEFL total score was used as the matching variable.)

Third, graduate-undergraduate differences were evaluated in light of differences in TWE performance across area of study. An examinee's performance on TWE may reflect the demands of the programs to which he or she is applying. Bridgeman and Carlson (1983) surveyed the ways in which undergraduate English faculty and graduate school faculty in six fields--business management (MBA), civil engineering, electrical engineering, psychology, chemistry, and computer science--characterized the writing requirements in their programs. The researchers found that faculty responses differed across areas of study. For example, they found that "skill in arguing for a particular position is seen as very important for undergraduates, MBA students, and psychology majors, but of very limited importance in engineering, computer science, and chemistry" (Bridgeman & Carlson, 1983, p. 55). A review of the literature on writing tasks in academic programs is provided by Hale, Taylor, Bridgeman, Carson, Kroll, and Kantor (1994), whose field research is still in progress.

Past analyses of data from administrations dating from September 1989 to May 1991 showed that graduate applicants in the humanities and social sciences had higher mean TWE scores than did applicants in the biological and physical sciences (*TOEFL Test of Written English Guide*, 1992, p. 17). The TOEFL database includes the intended field of study of graduate applicants who request that their scores be reported to institutions. Therefore, we were able to compare graduates from each of five broad areas of study to undergraduates and to compare graduate applicants across areas of study. The results of these field-of-study comparisons provided a baseline for interpreting the overall graduate-undergraduate differences by revealing how much variation in average TWE scores (conditional on TOEFL score) could be expected to result from differences across field.[1]

[1]Sample sizes did not allow other demographic factors, such as gender, native language, and native country, to be taken into account in the primary analyses. These factors were examined post hoc.

Previous Research on Group Differences in Essay Performance

Reviews of past research on group performance differences on essay and multiple-choice tests are provided by Mazzeo, Schmitt, and Bleistein (1992), who studied gender differences on constructed-response and multiple-choice components of four Advanced Placement (AP®) tests, and by Pomplun, Wright, Oleka, and Sudlow (1992), who conducted a study of essay performance on the College Board English Composition Test. As noted by Pomplun et al., one of the two most common approaches to the investigation of group differences in essay performance is *descriptive analysis*, which examines mean between-group differences (possibly expressed in standard deviation units) on the multiple-choice component and on each available essay item (e.g., the AP analyses by Morgan, 1992 and Morgan & Maneckshana, 1992, and certain TOEFL/TWE analyses by Golub-Smith et al., 1993). The other frequently used approach is *conditional analysis*, which typically compares essay performance for each group, conditional on the score on an accompanying multiple-choice test. The purpose of conditional analysis is to compare item performance for group members who are similarly proficient in the area of interest. An example of the conditional approach is the use of linear regression to predict essay performance using multiple-choice scores within each group of interest, followed by a comparison of the estimated regression lines across groups, as in Pomplun et al. (1992).

A conditional approach that has fewer assumptions than linear regression, and has been applied successfully by Zwick, Donoghue, and Grima (1993a; 1993b), is the test of Mantel (1963). An overview and example of the analysis approach are given in subsequent sections. Details of the procedure, including formulas, appear in Zwick and Thayer (1994; in press) and Zwick, Donoghue, and Grima (1993a; 1993b).

Methods

Mantel's Statistical Procedure for Ordered Variables

The statistical procedure developed by Mantel (1963) can be used to compare essay means for two groups, conditional on a matching variable. In this study, the Mantel procedure was used to test whether performance on TWE differed to a statistically significant degree for graduate and undergraduate examinees who had performed similarly on TOEFL and to conduct related group comparisons.

The Mantel statistic has been applied for research purposes to test items from the Graduate Record Examinations music test by Holland and Thayer (Holland, 1991), to essay items from the National Assessment of Educational Progress (Allen and Donoghue, 1994; Zwick, Donoghue, and Grima, 1993a; 1993b), and to simulation data (Allen and Donoghue, 1994; Chang, Mazzeo, and Roussos, 1995; Mazzeo and Chang, 1994; Welch and Hoover, 1993; Zwick, Donoghue, and Grima, 1993a; 1993b; Zwick and Thayer, 1994; in press). The results of Zwick, Donoghue, and Grima (1993a; 1993b) support the utility of the method for investigating group differences in performance on measures like TWE.

Descriptive Index of Differential Performance

The Mantel procedure indicates only whether a difference between matched groups is statistically significant. Even if a difference in average TWE scores between matched graduate and undergraduate applicants was found to be statistically significant, it might be small and not of practical importance. A measure of the *size* of the difference, expressed in terms of TWE score units, can be a useful supplement to the significance tests. A descriptive index that is applicable to items that are scored on an ordered scale was proposed by Dorans and Schmitt (1991). The index compares the item means of two groups, adjusted for differences in the distribution of members of the two groups across the values of the matching variable. (The statistic is an extension of the standardization statistic developed by Dorans & Kulick, 1986, for summarizing differential item functioning in the case of dichotomous items.) In this report, the statistic is called the standardized mean difference, or SMD. Two possible standard error formulas for SMD were derived by Zwick (1992; Zwick and Thayer, 1994; in press).[2] The calculation of SMD and an accompanying standard error may help the user interpret the results of Mantel's approach.[3]

---

[2]The standard error of a statistic is a measure of how precisely that statistic is estimated. The greater the precision, the smaller the standard error.

[3]Recent work by Mazzeo and Chang (1994) and Chang, Mazzeo, and Roussos (1995) showed that, under some conditions, the Mantel and SMD procedures do not perform as well as a competing method--an extension to polytomous items of the SIBTEST procedure (Shealy & Stout, 1993). For several reasons, we proceeded to use the Mantel and SMD methods. Possible problems with these approaches are minimized when (1) the mean difference between the two groups on the matching variable is not large and (2) the matching variable is very reliable (see Zwick & Thayer, 1994; in press). Both these conditions obtain in the present study. Also, an operational version of the SIBTEST program for polytomous items was not available. Finally, the Mantel and SMD procedures are more familiar and easier to understand.

Hypothetical Example

Suppose that the responses of graduate and undergraduate examinees to a single TWE prompt are as shown in the top two panels of Table 1. Although TWE is scored on a 1-6 scale, with half-point intervals, assume, for simplicity of presentation, that the essay is scored on a 1-3 scale. The top two panels of Table 1 represent two levels of the TOEFL matching variable. (In an actual analysis there would, of course, be many more levels of the matching variable.) The numbers in the body of the top two panels of Table 1 represent frequencies of examinees. For example, the "13" in the upper left of the top panel indicates that, among low scorers on the TOEFL matching variable, 13 graduate applicants received an essay score of "1."

The (unadjusted) difference between the essay means for the two groups, obtained by subtracting the graduate examinee mean (2.31) from the undergraduate examinee mean (2.22), was -.09. This value is shown at the foot of Table 1, along with other summary statistics. Simply comparing these essay means would lead to the conclusion that undergraduates did not perform as well on this prompt as graduates. Although the simple mean difference was negative, note that the Mantel Z value (see Zwick and Thayer, 1994; in press) was positive (1.63), reflecting the fact that *within* each level of the matching variable, the undergraduate group had a higher TWE mean than the graduate group, as shown in the summary panel of Table 1. The negative (unadjusted) mean difference occurred because the graduate group members were more likely than the undergraduate group members to receive a high score on the TOEFL matching variable, and high scores on the matching variable were associated with high TWE scores. The SMD adjusts for the differences in the distribution of graduates and undergraduates across levels of the matching variable by "standardizing" the means for both groups of examinees and subtracting the standardized mean for graduates (2.26) from the standardized undergraduate mean (2.54). The standardization was achieved by weighting the "Low" and "High" tables by the proportion of examinees in that table (48/241 and 193/241, respectively). The resulting value of SMD was positive (.28), like the Mantel Z statistic, with a standard error of 0.17.[4] This SMD value indicates that the between-group difference in mean item score (undergraduate - graduate) was 0.28 of

[4]The weighting scheme used in this example and throughout the present study is not identical to that used in Zwick and Thayer (1994; in press). The approach used here is discussed in Dorans and Kulick (1986). To compute the standard error of SMD, the hypergeometric formula in Zwick and Thayer (1994; in press) was used (with the revised weights).

a score point, after adjusting for group differences in the distribution of the matching variable.

Based on the Mantel and SMD statistics, the conclusion would be that, after matching examinees on a measure of overall language proficiency, undergraduates performed better than graduates, rather than worse, although the difference was not statistically significant. The SMD statistic, like the Mantel Z statistic, reflects the superior performance of undergraduates *within* each level of the matching variable. This illustration shows how *conditional* analysis can produce different and more interpretable results than simple comparisons of mean TWE scores.

TWE Analyses

TWE Data Sets

Our analyses covered 22 TWE administrations between October 1988 and October 1993 (see Table 2).[5] A separate analysis was conducted for each of three geographical regions, yielding 66 sets of results.[6] Each analysis was based on a single TWE prompt.

The three regions can be roughly characterized as follows:

Region 1: Asia and the Pacific

Region 2: Africa, the Middle East, and Europe

Region 3: North, Central, and South America

---

[5]The October 1989 data set was excluded because it was part of a special study (Golub-Smith et al., 1993) that involved the administration of eight different prompts within each region. The data sets that were available to us did not permit us to identify the prompt each examinee received.

[6]In most cases, a different prompt had been administered to the three regions; in some cases, the same prompt had been used. (Security considerations prevent the inclusion of further detail about the assignment of prompts and the definition of regions.) Even in administrations that used the same prompt across regions, combining data across regions would not have been advisable because it could have produced an undesirable confounding of region effects with effects of graduate/undergraduate status.

However, the definition of the regions varies to some degree over administrations, and some countries do not participate in every administration.

Editing. The data files provided to us included records for all individuals who registered for TOEFL or the Test of Spoken English (TSE®). Preparing the data for analysis required the following steps:

1. Data files that shared a common format were obtained. TOEFL archival files were available for administrations that took place between October 1988 and October 1991. More recent data files had to be converted to the same format as the archival files.

2. A data file was created for each of the 22 administrations, and a region code was added. This required that a computer program be written to determine each examinee's region by using test center codes. The plausibility of the resulting counts of examinee records was checked by comparing them to the counts in TWE test analysis reports.

3. In each data file, records unsuitable for our analyses were eliminated. Specifically, records were deleted for candidates who requested that their data not be used for research, candidates whose status was neither "undergraduate" nor "graduate," and candidates with invalid TWE or TOEFL scores. Table 2 shows the initial number of candidate records and the number of usable TWE records for each of the 22 included administrations. Also given are the percentages of records excluded because the candidate (1) did not have a valid TWE score, (2) did not have a valid TOEFL score, or (3) did not indicate that he or she was applying to a graduate or undergraduate institution. The number of records excluded for all other reasons was very small. The percentages shown for Reason 1 include candidates who registered but did not appear on the day of testing and candidates who registered for the Test of Spoken English (TSE®) only. (Percentages were computed in a hierarchical fashion; i.e., candidates excluded for Reason 1 were not included in the percentages excluded for Reasons 2 and 3, even if these reasons applied.) The number of usable records for a TWE administration ranged from about 16,000 to about 81,000.

Preliminary Analyses

In our initial descriptive analyses of the TWE data, we examined the TWE and TOEFL score distributions, the TWE nonresponse rates, and the correlations between the scores assigned by the two readers. Descriptive statistics were

computed separately for graduate and undergraduate applicants. A portion of these analyses is summarized in Tables 3-5.

TWE Scores. Every TWE response is read by two readers, each of whom assigns a holistic score ranging from 1 to 6. If the difference between the two readers' scores is less than two, the final TWE score is the average of the two readers' scores; thus, 11 distinct scores are possible. If the difference between the readers' scores is two points or greater, a chief reader resolves the discrepancy (see *TOEFL Test of Written English Guide*, 1992, p. 8). As shown in Tables 3-5, the percentage of cases in which a chief reader intervened was rarely more than 2 percent in Regions 1 and 3; it was usually between 2 percent and 5 percent in Region 2. Tables 3-5 also give the Pearson correlation between the scores of the two readers. These values ranged from .69 to .85.[7] The correlations between TWE and TOEFL scores (with no corrections applied) are also tabled; these ranged from .54 to .75. The inter-reader correlations and TWE-TOEFL correlations tended to be higher for undergraduates than graduates. Mean TWE scores are also shown. These were higher for graduates in two-thirds of the 66 data sets. For both graduates and undergraduates, the median standard deviation (not shown) of TWE across the 22 administrations was about 0.9 for Regions 1 and 2 and 0.8 for Region 3.

The percentage of TWE records with a "no-response" code is also tabled. These records are given a TWE score of 1 in the TOEFL data files and therefore had not been excluded from these preliminary analyses at the editing phase.[8] They were, however, excluded from all the group comparisons described in subsequent sections. (Records with an "off-topic" response are not given a TWE score and therefore were excluded from both the preliminary analyses and the group comparisons.)

TOEFL Scores. The TOEFL consists of 150 items divided among three separately scored sections--Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. The TOEFL total score is obtained by summing and rescaling the section scores. The total score can range from 200 to 677 (*TOEFL Test and Score Manual*, 1992, p. 16). Based on test

---

[7]Note that this correlation underestimates the reader reliability, since the assigned TWE score is the mean of the reader scores. The Spearman-Brown prophecy formula could be used to obtain an estimate of the reader reliability from these correlations, as in Golub-Smith et al. (1993).

[8]For purposes of score reporting, a "no-response" indicator from a separate field is used to distinguish these records from those of examinees who did respond and received a score of 1.

forms administered between July 1989 and June 1991, the median reliability for the TOEFL total score was found to be .95 (*TOEFL Test and Score Manual*, 1992, p. 31). Mean TOEFL scores for each administration are given in Tables 3-5.[9] The means were higher for graduates than for undergraduates in all 66 data sets. For undergraduates, the median standard deviation (not shown) of TOEFL across the 22 administrations was about 70 in all three regions; for graduates, the region medians were slightly smaller.

Group Comparisons

The Mantel statistical significance tests were performed and the SMD measures of the size of the standardized graduate-undergraduate difference were computed for each of the 66 data sets. Separate comparisons to undergraduates were also conducted for graduate applicants in each of five fields who requested score reports (and therefore stated their intended fields of study). Also, graduate applicants from each field of study were compared to each other. Finally, some supplementary analyses were conducted to explore the possible confounding effects of gender, native country, and native language. To assure stability of results, group comparisons were conducted only if at least 200 examinees in each group were available for analysis. Each of these types of analysis is discussed in subsequent sections.

Choice of TOEFL Total Score as a Matching Variable. For several reasons, the TOEFL total score appeared to be the best choice for a matching variable.[10] First, the TOEFL total score is more reliable than any of the section scores (*TOEFL Test and Score Manual*, 1992, p. 31). Second, it is typically more highly correlated with TWE than any of the section scores, even after correcting for unreliability

---

[9]In the data we analyzed, it was generally the case that, within a given administration and region, the same TOEFL form was used. In five of the 66 data sets, however, exceptions to this rule occurred, with the result that the matching variable was not the same for all individuals. However, because TOEFL forms are equated, scores should have approximately the same meaning across forms.

[10]In the case of dichotomous items, Mantel's statistic reduces to the Mantel-Haenszel (MH; 1959) test, which is the basis for the DIF procedure of Holland and Thayer (1988). Research on the MH test has shown that the studied item should be included in the matching variable (Holland & Thayer, 1988; Zwick, 1990; Donoghue, Holland & Thayer, 1993). Zwick, Donoghue, and Grima (1993a; 1993b) showed that this finding can be generalized to items that are scored on an ordered scale as well. The matching variable here, therefore, was actually the sum of the TOEFL total score and the score on the studied TWE item. Examinees who fell within the same five-point interval (i.e., 201-205, 206-210, etc.) were considered to be matched.

(*TOEFL Test of Written English Guide*, 1992, p. 13). Finally, its correlation with TWE does not appear to vary systematically across geographical region, unlike the correlation of TWE with scores on TOEFL Sections 1 and 2. The median correlation of the TOEFL total score with TWE for the eight administrations tabled in the *TOEFL Test of Written English Guide* (1992, p. 13) is .66 for Regions 1 and 2 and .65 for Region 3. (The *Guide* notes that these values have been corrected for unreliability of TOEFL.) These results are consistent with the TOEFL-TWE correlations for the data used in the present study (Tables 3-5). The statistical procedures that were used to compare groups do not require that the matching variable measure exactly the same skills as TWE.

Graduate-undergraduate Comparisons. Tables 6-8 give the results of the graduate-undergraduate comparisons for Regions 1, 2, and 3, respectively. Additional summary information is given in the top line of Tables 9-11, which summarize the graduate-undergraduate comparisons, as well as the field-of study comparisons described below. The first column of Tables 6-8 gives the value obtained by subtracting the graduate mean from the undergraduate mean. In 44 of the 66 comparisons, this simple mean difference was negative. In 41 of these 44 cases, however, the SMD was positive; that is, using a conditional analysis showed superior undergraduate performance, though a simple comparison of means had shown superior graduate performance. In all three regions, the median across the 22 administrations of the simple mean differences was slightly negative: -0.02, -0.03, and -0.04, for Regions 1, 2, and 3, respectively. By contrast, the standardized mean differences (SMDs) had medians of 0.18, 0.05, and 0.10, respectively. The 25th percentiles of the across-administration SMD distributions were 0.14, 0.03, and 0.08 for the three regions; the 75th percentiles were 0.21, 0.09, and 0.10. In 17 of the 22 administrations, Region 1 had the largest SMD followed by Region 3 and then Region 2.

The SMD values, like the simple mean differences, are in the TWE score-point metric and do not depend on sample size. There are several ways to evaluate the magnitude of the SMDs. Tables 9-11 show that the standard errors of the SMD statistics were typically between 0.01 and 0.02; therefore, the SMDs tended to be large relative to their standard errors. Application of the Mantel procedure (using a two-sided test at $\alpha = .01$) yielded statistically significant results in 60 of 66 comparisons. In all but two of these, the SMD was positive, indicating superior performance for undergraduates, conditional on TOEFL.[11] However, when sample sizes are large, as they are for these comparisons, very small differences that are of

[11] An alternative statistic can be obtained by dividing SMD by its standard error. The results are nearly identical to those obtained using Mantel's test.

little practical importance may be statistically significant. It is useful, therefore, to apply some other method of evaluating the findings. We therefore considered the size of the SMD values, focusing on those that exceeded 0.2 in magnitude. Because the TWE standard deviations averaged between 0.8 and 0.9, an SMD of 0.2 represents roughly one-fifth to one-fourth of a standard deviation unit. In considering mean differences, Cohen (1988) regards one-fifth of a standard deviation as a small effect. Although the present application involves conditional, rather than simple mean differences, Cohen's rule of thumb may still be useful as an approximate guideline.

The top line of Tables 9-11 shows the number of SMD values that exceeded 0.2 in magnitude for the graduate-undergraduate comparisons. In Region 1, there were eight such cases (October 1988, May 1989, September 1989, March 1990, September 1990, October 1991, March 1992, and May 1992); in Region 2, there was one (October 1993); and in Region 3, there were none. It is striking that eight of the nine SMDs exceeding 0.2 were concentrated in Region 1 between October 1988 and May 1992. Two of these SMDs occurred in administrations in which all three regions received the same essay prompt. Regions 2 and 3 had SMDs that did not exceed 0.1 for these administrations.

Another way to evaluate the SMDs for the graduate-undergraduate analyses is to compare them to the SMDs obtained from comparisons of graduates from different fields of study. In general, these field-of-study comparisons did not produce consistent patterns of SMDs exceeding 0.2, such as that observed in the Region 1 graduate-undergraduate analysis. A notable exception was the comparison of examinees in the social sciences to those in the physical sciences in Region 2, which showed substantial evidence of superior performance by social science examinees. The field-of study analyses are discussed in detail in the next section.

Field-of-study Comparisons. Field of study was categorized as in TOEFL's candidate bulletin: biological sciences, physical sciences, social sciences, humanities, law, and business. (Public health, which is listed under both social sciences and biological sciences in the candidate bulletin, was included in biological sciences.) The law group was too small to include in the analyses, leaving five fields. The analyses based on field of study had some substantial limitations. First, intended field of study is not available for undergraduates. Second, less than half of the graduate applicants included in our data had provided information about intended field of study. For the 66 TWE data sets, the percentages of graduates for whom field of study was missing ranged from 33 to 87. The percentage exceeded 50 in all but two of the data sets.

Two types of field-of-study analyses were conducted. The first compared (the pool of) undergraduates to graduates in each of five fields. The goal of these analyses was to provide information that could help to illuminate the finding that undergraduates perform better on TWE than graduates, conditional on TOEFL. The second type of analysis compared graduates across field of study. A separate analysis was conducted for each of the ten possible pairings of the five fields. The results of both types of field-of-study analysis are summarized in Tables 9, 10, and 11 for Regions 1, 2, and 3, respectively. The first column of figures gives the number of comparisons (out of a maximum of 22) for which sample sizes were adequate for analysis. Summary statistics are given for the simple mean difference, SMD, standard error of SMD, and Mantel Z value.

Considering the results for all three regions, several trends are evident. First, as noted earlier, the field-of-study comparisons produced results that were less consistent across administrations than the graduate-undergraduate comparisons. With the one exception noted earlier, the field-of-study comparisons were also less likely to yield SMDs with magnitudes exceeding 0.2. In general, the comparisons suggested that, for groups matched on the TOEFL score, applicants to graduate programs in the physical and biological sciences tended to have lower TWE means than graduates in the social sciences and undergraduates. Graduate students in the physical sciences tended to have lower TWE means than matched graduates in the humanities. (This is largely consistent with the results in the *TWE Guide*, cited above, on TWE performance differences across field of study.) The comparison between undergraduates and graduates in the social sciences yielded contradictory results: in Region 1, the SMDs showed substantial evidence of superior performance by undergraduates, whereas in Region 2, the opposite was true.

Supplementary Analyses of the Effects of Gender, Native Country, and Native Language. Interpretation of the graduate-undergraduate comparisons is not straightforward because graduate/undergraduate status is confounded with many other factors. The distributions of examinees across such variables as field of study, gender, native country, and native language may differ substantially for graduates and undergraduates. For example, graduates may be more likely to be concentrated in the sciences than undergraduates, and this disparity may be greater in Region 1. (Because information about intended field of study is unavailable for undergraduates and for most graduates, this hypothesis cannot be tested with existing data.)

The possible confounding effects of gender were also considered. Golub-Smith et al. (1993) compared men and women within eight spiral samples of examinees, each of which received a different TWE essay prompt. In all eight

instances, the women had a lower average TOEFL score, but a higher average TWE score. Research on other tests, such as the Advanced Placement examinations, has often found that relative to men, women performed better on essay items than on multiple-choice items in the same subject area (Bridgeman & Lewis, 1994; Mazzeo, Schmitt & Bleistein, 1992; Morgan, 1992; Morgan & Maneckshana, 1992). Male-female performance differences also tend to vary across essay topics. In the present study, we considered the possibility that the superior performance of undergraduates, relative to matched graduates, might be attributable to superior performance of women on TWE (conditional on TOEFL) combined with the higher proportion of women among undergraduates than among graduates. This hypothesis seemed more plausible because the graduate-undergraduate disparity in the proportion of women tended to be largest in Region 1, where the largest SMDs were concentrated. We conducted some additional analyses to explore this issue. For example, in select administrations, we compared graduates and undergraduates separately for male and female examinees. We found that, for females, the graduate-undergraduate SMD tended to be smaller than in the combined-sex analysis, but for males, it tended to be larger. These results suggest that the graduate-undergraduate findings cannot be explained by gender alone.

Another hypothesis we explored was that the pattern of graduate-undergraduate results could be explained by cultural and language differences between graduates and undergraduates. In 10 administrations, we examined the native country and native language of graduates and undergraduates in all three regions. While country and language differences between graduates and undergraduates did tend to be greater in Region 1 than in the other two regions, we could not find any factor that was consistently related to the occurrence of large SMDs. In some cases, large demographic differences were associated with small SMDs; in other cases, the reverse was true. It also became apparent that, especially for Region 1, the composition of the region changed significantly across administrations. Two major reasons for this inconsistency are that the definition of the regions varies over time and that some countries participate in only a subset of the administrations.

Summary and Discussion

The performance of graduate and undergraduate school applicants on the Test of Written English (TWE) was compared for each of 66 TWE data sets (22 administrations *X* 3 regions), dating from October 1988 to October 1993. Specifically, the analyses compared the average TWE score for graduates and undergraduates after matching examinees on the TOEFL total score. In evaluating the magnitude of the graduate-undergraduate differences, the effects of other key variables, such as field of graduate study, gender, native country, and native language, were investigated. The main findings were as follows:

1. Undergraduates tended to perform better on TWE than graduates with similar scores on TOEFL. For graduates and undergraduates who were matched on the TOEFL score, undergraduate TWE means were higher than graduate means in 63 of the 66 analyses. Thus, the direction of these differences was consistent over a wide variety of TWE topics. Although most were statistically significant, the standardized mean differences never exceeded 0.3 of a TWE score point. They are noteworthy, however, because they give a different picture from that obtained by simply comparing means for graduates and undergraduates (without matching): in 44 of 66 such comparisons, mean TWE scores were higher for *graduate* applicants. The largest difference was about 0.3 of a TWE score point, in favor of graduates.

2. Of the nine standardized mean differences (out of 66) that exceeded 0.2, eight were in Region 1 between October 1988 and May 1992.

3. Field-of-study comparisons suggested that, for groups matched on the TOEFL score, applicants to graduate programs in the physical and biological sciences tended to have lower TWE means than undergraduates and graduates in the social sciences. Graduate students in the physical sciences tended to have lower TWE means than matched graduates in the humanities. This is largely consistent with results appearing in the *TWE Guide* for (unmatched) comparisons across field of study.

Interpretation of the graduate-undergraduate comparisons is complicated because the distributions of examinees across such variables as field of study, gender, native country, and native language sometimes differ substantially for graduates and undergraduates. We could not find any strong evidence that differences in the distribution of gender, native language, and native country were associated with the size of the graduate-undergraduate SMDs, although demographic

differences between graduates and undergraduates did appear to be more prominent in Region 1, where most of the larger SMDs occurred. Results of comparisons based on field of study suggest the possibility that graduate-undergraduate differences may result in part from a greater concentration of graduate applicants in scientific fields, but existing data do not allow exploration of this hypothesis.

The hypothesis advanced most often by TOEFL and TWE staff about the preponderance of SMDs favoring undergraduates was that undergraduates are more likely than graduates to have recently participated in intense English writing instruction. In contrast, graduates are more likely to have focused their recent studies on their chosen fields of specialization. It is possible that this curricular difference is more substantial within Region 1. Again, however, this hypothesis cannot be tested with existing data.

In addition to assessing the relevance of demographic and instructional factors, we considered hypotheses about the prompts themselves. We examined the topics of the TWE prompts and their explicitness (see Golub-Smith et al., 1993), but were unable to identify any characteristics that appeared to be related to the pattern of performance differences. We also solicited the views of TOEFL and TWE personnel, including test development staff, program directors, and committee members, about the relevance of the prompts to our findings. Again, no relevant attributes of the prompts were identified. In any case, the *direction* of the conditional graduate-undergraduate differences was apparently unrelated to characteristics of the prompts.

A study like this one cannot, in itself, demonstrate whether different forms of TWE are needed for graduate and undergraduate applicants. Our research primarily serves to reveal existing differences in TWE *performance* between undergraduates and graduates and to produce hypotheses about reasons for these differences. If an exploration of the influence of instructional experience and field of study on TWE performance is deemed to be of interest, the TOEFL program could collect information from examinees on these variables. To further examine the question of separate TWE forms, it might be fruitful to conduct a validity study that investigated the degree to which TWE scores predict the writing performance of graduates and undergraduates in academic settings. At the least, determining whether there is a need for separate TWE forms necessitates an analysis of writing demands in the undergraduate and graduate programs requiring TWE, such as the ongoing investigation by Hale, Taylor, Bridgeman, Carson, Kroll, and Kantor (1994).

In light of the variation in writing demands across academic fields (e.g., Bridgeman & Carlson, 1983; Hale et al., 1994) and the corresponding performance differences observed in the present study, the development of TWE prompts that are tailored to particular fields of study also could be considered. The utility of this idea, too, could be explored in the context of a validity study, which could be designed to include separate analyses within selected academic fields.

References

Allen, N. L., & Donoghue, J. R. (1994). *DIF analysis based on complex samples of dichotomous and polytomous items.* Presented at the annual meeting of the American Educational Research Association, New Orleans.

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students.* TOEFL report 15/ ETS Research Report 83-18. Princeton, New Jersey: Educational Testing Service.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37-50.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communication, 8*, 533-556.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure.* ETS Research Report 95-5. Princeton, New Jersey: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, New Jersey: Erlbaum.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (eds.) *Differential Item Functioning.* Hillsdale, New Jersey: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach.* ETS Research Report 91-47. Princeton, New Jersey: Educational Testing Service.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English.* TOEFL Report 42/ ETS Research Report 93-10. Princeton, New Jersey: Educational Testing Service.

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (April 1994). *A study of writing tasks assigned in academic degree programs.* Draft final report submitted to the TOEFL Research Committee.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom.* Cambridge, England: Cambridge University Press.

Holland, P. W. (January 14, 1991). *Item and DIF analyses for items with ordered responses.* Internal ETS memorandum.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity,* pp. 129-145. Hillsdale, New Jersey: Erlbaum.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mazzeo, J., & Chang, H.-H. (April 1994). *Detecting DIF for polytomously scored items: progress in adaptation of Shealy-Stout's SIBTEST procedure.* Presented at the annual meeting of the American Educational Research Association, New Orleans.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1992). *Sex-related performance differences on constructed response and multiple-choice sections of Advanced Placement Examinations.* College Board Report No. 92-7. New York: College Entrance Examination Board.

Morgan, R. (May 20, 1992). *Subgroup performance of AP free response items.* ETS memorandum.

Morgan, R., & Maneckshana, B. (March 5, 1992). Subgroup reliability for the Advanced Placement English Language and Composition, European History, and United States Government and Politics Examinations (Form 3NBP). ETS memorandum.

Pomplun, M., Wright, D., Oleka, N., & Sudlow, M. (1992). *An analysis of English Composition essay prompts for differential difficulty.* College Board Report 92-4, ETS Research Report 92-34. Princeton, New Jersey: Educational Testing Service.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

TOEFL Test of Written English Guide (third edition). (1992). Princeton, New Jersey: Educational Testing Service.

TOEFL Test and Score Manual (1992-1993 edition). (1992). Princeton, New Jersey: Educational Testing Service.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1-19.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185-197.

Zwick, R. (1992). *Application of Mantel's score test to the analysis of differential item functioning for ordinal items.* Internal ETS technical report.

Zwick, R., Donoghue, J. R., & Grima, A. (1993a). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). *Assessing differential item functioning in performance tests.* ETS Research Report 93-14. Princeton, New Jersey: Educational Testing Service.

Zwick, R., & Thayer, D. T. (1994). *Evaluation of the magnitude of differential item functioning in polytomous items.* ETS Research Report 94-13. Princeton, New Jersey: Educational Testing Service.

Zwick, R., & Thayer, D. T. (in press). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics.*

Table 1

Hypothetical Example of the Mantel Approach and SMD Statistic:
Frequencies of Graduates (G) and Undergraduates (UG) Receiving Each TWE Score

Low Score on TOEFL Matching Variable

	Essay Score			
Group	1	2	3	Total
Graduates	13	5	7	25
Undergraduates	5	14	4	23
Total	18	19	11	48

High Score on TOEFL Matching Variable

	Essay Score			
Group	1	2	3	Total
Graduates	28	54	98	180
Undergraduates	1	2	10	13
Total	29	56	108	193

Summary of Results

	Low on TOEFL Matching Variable			High on TOEFL Matching Variable			Combined High and Low		
Statistic	G	UG	Total	G	UG	Total	G	UG	Total
Proportion of cases	.12	.64	.20	.88	.36	.80	1.00	1.00	1.00
TWE mean	1.76	1.96	1.85	2.39	2.69	2.41	2.31	2.22	2.30
Standardized mean	-	-	-	-	-	-	2.26	2.54	-

UG mean minus G mean = 2.22 - 2.31 = -0.09

Mantel Z = 1.63

SMD = 2.54 - 2.26 = 0.28

Standard error of SMD = 0.17

Table 2
Examinee Data Included in the Study

Administration	Number of Candidate Records	Percentage Excluded[a]			Number of Usable TWE Records
		1	2	3	
October, 1988	74,999	13	1	9	57,536
March, 1989	70,537	17	2	12	48,528
May, 1989	95,639	37	2	8	67,288
September, 1989	24,814	20	0	14	16,357
March, 1990	82,247	19	0	13	55,004
May, 1990	113,568	16	0	12	80,365
September, 1990	28,442	22	0	14	18,328
October, 1990	102,204	15	0	11	76,086
March , 1991	91,531	18	0	13	62,009
May, 1991	114,997	17	0	12	81,462
September, 1991	36,033	19	0	12	24,744
October, 1991	113,577	19	0	9	80,285
March , 1992	96,817	18	0	13	67,067
May, 1992	114,392	16	0	12	81,489
August, 1992	78,389	16	0	11	57,731
September, 1992	30,235	19	0	13	20,683
October, 1992	110,192	28	0	9	68,933
February, 1993	72,479	18	0	13	50,168
May, 1993	106,725	16	0	13	75,681
August, 1993	71,201	17	0	12	50,856
September, 1993	36,069	18	0	13	24,863
October, 1993	108,758	15	0	10	81,046
TOTAL	1,803,960				1,246,509

[a] Reasons for exclusion are as follows:

1. No valid TWE score (includes registrants for the Test of Spoken English and registrants who did not appear for testing)
2. No valid TOEFL score
3. Examinee status is neither undergraduate nor graduate

Percentages excluded were computed in a hierarchical fashion; i.e., an examinee who has been excluded for Reason 1 will not be included in the exclusion percents for Reasons 2 and 3, even if these reasons also apply. Percentages have been rounded to the nearest integer.

Table 3
Preliminary Analysis - Region 1

	Reader Correlation		Pct.[a] No Response		Pct.[a] Chief Reader		Mean TWE		Mean TOEFL		TWE-TOEFL Correlation	
Administration	G	UG	G	UG	G	UG	G	UG	G	UG	G	UG
October, 1988	0.78	0.78	1	1	1	2	3.72	3.63	544	510	0.65	0.72
March, 1989	0.79	0.78	1	0	1	1	3.59	3.62	522	507	0.66	0.72
May, 1989	0.75	0.82	1	1	1	1	3.73	3.77	529	509	0.62	0.75
September, 1989	0.79	0.81	1	1	0	0	3.71	3.60	521	478	0.54	0.63
March, 1990	0.80	0.80	1	1	1	1	3.36	3.44	514	500	0.64	0.72
May, 1990	0.78	0.81	2	2	1	2	3.69	3.70	531	511	0.60	0.70
September, 1990	0.83	0.84	2	2	0	0	3.59	3.48	526	481	0.57	0.65
October, 1990	0.81	0.82	2	2	0	1	3.74	3.71	534	510	0.68	0.73
March, 1991	0.84	0.83	1	1	1	2	3.76	3.67	521	500	0.65	0.71
May, 1991	0.82	0.85	3	3	0	1	3.56	3.54	534	512	0.59	0.69
September, 1991	0.81	0.82	2	1	1	1	3.80	3.65	527	491	0.60	0.66
October, 1991	0.78	0.83	5	4	1	2	3.67	3.70	539	511	0.56	0.67
March, 1992	0.80	0.82	3	1	1	2	3.56	3.72	516	503	0.55	0.68
May, 1992	0.78	0.82	6	5	0	1	3.60	3.63	527	505	0.58	0.68
August, 1992	0.78	0.80	4	4	1	1	3.73	3.63	542	514	0.59	0.65
September, 1992	0.82	0.83	4	3	0	0	3.79	3.45	531	494	0.63	0.67
October, 1992	0.80	0.78	2	1	1	1	3.83	3.67	544	511	0.67	0.70
February, 1993	0.83	0.82	3	3	1	4	3.62	3.64	529	511	0.65	0.68
May, 1993	0.75	0.78	5	4	0	1	3.70	3.65	542	513	0.59	0.68
August, 1993	0.77	0.80	4	3	0	1	3.74	3.77	545	522	0.62	0.67
September, 1993	0.84	0.79	6	3	0	1	3.78	3.90	537	526	0.66	0.65
October, 1993	0.78	0.79	4	3	1	1	3.77	3.58	554	514	0.63	0.66

[a] Percentages have been rounded to the nearest integer

Table 4
Preliminary Analysis - Region 2

	Reader Correlation		Pct.[a] No Response		Pct.[a] Chief Reader		Mean TWE		Mean TOEFL		TWE-TOEFL Correlation	
Administration	G	UG	G	UG	G	UG	G	UG	G	UG	G	UG
October, 1988	0.76	0.75	2	1	5	4	3.95	3.80	558	534	0.63	0.66
March, 1989	0.71	0.69	4	2	4	5	3.96	3.97	548	536	0.63	0.64
May, 1989	0.76	0.75	3	2	4	5	4.03	4.02	540	530	0.63	0.66
September, 1989	0.80	0.85	6	2	0	0	3.95	3.87	522	510	0.70	0.69
March, 1990	0.78	0.77	3	2	3	2	4.10	4.07	545	535	0.65	0.67
May, 1990	0.81	0.80	3	1	3	4	4.14	4.08	543	530	0.64	0.68
September, 1990	0.82	0.79	5	3	1	3	3.95	3.98	540	528	0.63	0.61
October, 1990	0.82	0.81	4	1	2	2	4.34	4.35	564	554	0.62	0.64
March, 1991	0.82	0.82	5	2	3	4	4.03	4.02	554	545	0.55	0.56
May, 1991	0.76	0.76	6	3	2	2	3.95	3.94	555	544	0.57	0.60
September, 1991	0.75	0.79	5	3	4	4	4.09	3.97	548	528	0.64	0.66
October, 1991	0.78	0.78	3	3	2	6	4.20	3.87	572	542	0.64	0.67
March, 1992	0.77	0.78	5	4	1	3	4.01	3.88	561	547	0.58	0.62
May, 1992	0.80	0.81	5	4	1	2	4.08	3.96	546	538	0.62	0.65
August, 1992	0.78	0.80	11	6	8	9	3.77	3.83	539	532	0.64	0.67
September, 1992	0.82	0.85	11	8	2	2	3.96	3.81	549	527	0.69	0.70
October, 1992	0.78	0.79	8	5	2	3	4.04	4.00	565	550	0.65	0.62
February, 1993	0.79	0.81	10	8	2	3	3.78	3.77	542	530	0.63	0.68
May, 1993	0.76	0.77	12	7	2	2	3.92	3.91	550	544	0.61	0.64
August, 1993	0.76	0.76	6	3	2	1	3.64	3.46	537	504	0.64	0.67
September, 1993	0.78	0.78	6	4	0	0	3.71	3.50	534	498	0.62	0.61
October, 1993	0.80	0.80	5	5	2	3	3.69	3.87	546	542	0.67	0.66

[a] Percentages have been rounded to the nearest integer.

Table 5
Preliminary Analysis - Region 3

	Reader Correlation		Pct.[a] No Response		Pct.[a] Chief Reader		Mean TWE		Mean TOEFL		TWE-TOEFL Correlation	
Administration	G	UG	G	UG	G	UG	G	UG	G	UG	G	UG
October, 1988	0.74	0.75	0	0	3	2	3.70	3.66	525	507	0.63	0.68
March, 1989	0.73	0.72	1	0	2	2	3.68	3.66	521	504	0.63	0.65
May, 1989	0.72	0.73	1	0	3	2	3.77	3.79	512	505	0.62	0.66
September, 1989	0.76	0.75	1	0	1	1	3.88	3.77	512	486	0.58	0.62
March, 1990	0.78	0.79	1	1	2	2	3.70	3.64	520	506	0.65	0.67
May, 1990	0.78	0.77	1	1	2	2	3.95	3.94	519	508	0.62	0.66
September, 1990	0.78	0.79	3	1	1	1	3.76	3.69	522	497	0.58	0.62
October, 1990	0.79	0.80	2	1	1	1	3.65	3.64	527	571	0.61	0.66
March, 1991	0.80	0.81	2	1	1	1	3.81	3.81	520	504	0.59	0.62
May, 1991	0.76	0.77	2	1	1	1	3.73	3.76	520	509	0.59	0.64
September, 1991	0.74	0.75	4	3	1	1	3.79	3.70	519	492	0.58	0.62
October, 1991	0.76	0.75	3	2	2	2	3.82	3.78	525	508	0.63	0.67
March, 1992	0.75	0.76	4	2	1	1	3.71	3.71	526	510	0.60	0.65
May, 1992	0.74	0.76	3	2	1	1	3.58	3.62	515	505	0.57	0.63
August, 1992	0.71	0.73	4	3	2	2	3.71	3.62	527	504	0.58	0.62
September, 1992	0.79	0.81	7	5	1	0	3.71	3.59	524	501	0.62	0.63
October, 1992	0.75	0.77	5	3	1	1	3.86	3.80	530	509	0.58	0.62
February, 1993	0.72	0.73	5	3	1	1	3.66	3.67	521	504	0.58	0.62
May, 1993	0.73	0.76	6	4	1	1	3.69	3.71	525	516	0.58	0.63
August, 1993	0.77	0.79	6	4	1	1	3.72	3.65	528	507	0.59	0.63
September, 1993	0.74	0.77	6	3	0	0	3.78	3.69	529	505	0.55	0.59
October, 1993	0.75	0.76	6	3	1	1	3.72	3.67	534	511	0.62	0.65

[a] Percentages have been rounded to the nearest integer.

Table 6
Results of Graduate - Undergraduate Comparisons
Region 1

Administration	Mean Difference[a]	SMD[a]	S.E. of SMD	Mantel Z[b]	Sample Size UG	Sample Size G
October, 1988	-0.09	0.25	0.01	28.65	10,893	23,332
March, 1989	0.04	0.17	0.01	16.88	9,274	12,569
May, 1989	0.04	0.21	0.01	23.77	10,576	26,080
September, 1989	-0.11	0.22	0.02	13.86	4,158	5,060
March, 1990	0.08	0.20	0.01	22.55	11,147	12,988
May, 1990	0.01	0.18	0.01	23.60	13,707	28,557
September, 1990	-0.11	0.25	0.02	16.79	4,719	5,527
October, 1990	-0.03	0.19	0.01	28.69	15,080	27,498
March , 1991	-0.08	0.13	0.01	14.23	11,136	16,885
May, 1991	-0.02	0.16	0.01	19.43	12,829	28,775
September, 1991	-0.15	0.13	0.01	9.79	5,090	6,502
October, 1991	0.03	0.25	0.01	32.50	12,967	23,568
March , 1992	0.16	0.25	0.01	27.50	10,066	13,223
May, 1992	0.04	0.20	0.01	26.60	11,247	23,985
August, 1992	-0.10	0.14	0.01	17.20	11,849	25,167
September, 1992	-0.34	0.00	0.01	-0.23	4,556	5,342
October, 1992	-0.16	0.11	0.01	14.68	10,879	19,473
February, 1993	0.02	0.16	0.01	11.84	3,674	9,950
May, 1993	-0.05	0.18	0.01	21.87	10,481	24,706
August, 1993	0.03	0.19	0.01	21.95	6,970	18,255
September, 1993	0.12	0.19	0.02	10.35	1,905	4,011
October, 1993	-0.19	0.12	0.01	16.01	11,793	24,733
Median	-0.02	0.18	0.01			
Mean	-0.04	0.18	0.01			
S.D.	0.11	0.06	0.00			

[a] Positive values indicate that undergraduate performance was superior to graduate performance. Mean differences shown here may differ slightly from results obtained from the TWE means in Table 3 because Table 3 results were conducted before data editing and because of rounding.

[b] All p-values were less than .001, except for September 1992 (p=.816).

Table 7
Results of Graduate - Undergraduate Comparisons
Region 2

Administration	Mean Difference[a]	SMD[a]	S.E. of SMD	Mantel Z	p-value	Sample Size UG	Sample Size G
October, 1988	-0.15	0.04	0.02	2.43	0.015	3,331	3,706
March, 1989	0.01	0.09	0.01	6.72	0.000	4,055	4,183
May, 1989	-0.01	0.07	0.01	5.04	0.000	5,388	5,041
September, 1989	-0.06	0.02	0.04	0.69	0.489	806	636
March, 1990	-0.03	0.05	0.01	4.22	0.000	4,899	4,838
May, 1990	-0.06	0.05	0.01	3.81	0.000	6,469	6,169
September, 1990	0.04	0.12	0.04	3.42	0.001	796	779
October, 1990	0.02	0.10	0.02	6.34	0.000	3,520	4,432
March , 1991	-0.01	0.05	0.01	3.55	0.000	4,568	4,484
May, 1991	-0.01	0.06	0.01	5.33	0.000	6,711	5,842
September, 1991	-0.11	0.02	0.03	0.87	0.385	994	1,032
October, 1991	-0.32	-0.07	0.01	-5.56	0.000	4,808	10,824
March , 1992	-0.13	-0.02	0.01	-1.59	0.111	6,352	9,967
May, 1992	-0.12	-0.05	0.01	-4.47	0.000	7,275	9,689
August, 1992	0.07	0.13	0.02	6.11	0.000	1,786	2,158
September, 1992	-0.10	0.03	0.03	1.01	0.312	911	915
October, 1992	-0.04	0.08	0.01	6.14	0.000	4,385	5,334
February, 1993	-0.01	0.09	0.01	6.75	0.000	5,322	6,075
May, 1993	-0.01	0.03	0.01	2.81	0.005	5,801	5,458
August, 1993	-0.18	0.10	0.01	7.27	0.000	4,714	4,856
September, 1993	-0.21	0.04	0.01	2.83	0.005	4,345	4,490
October, 1993	0.18	0.20	0.01	17.48	0.000	4,861	9,679
Median	-0.03	0.05	0.01				
Mean	-0.06	0.06	0.02				
S.D.	0.10	0.06	0.01				

[a] Positive values indicate that undergraduate performance was superior to graduate performance. Mean differences shown here may differ slightly from results obtained from the TWE means in Table 4 because Table 4 results were conducted before data editing and because of rounding.

Table 8
Results of Graduate - Undergraduate Comparisons
Region 3

Administration	Mean Difference[a]	SMD[a]	S.E. of SMD	Mantel Z[b]	Sample Size UG	G
October, 1988	-0.04	0.10	0.01	9.65	8,991	6,807
March, 1989	-0.03	0.12	0.01	12.20	10,545	7,431
May, 1989	0.02	0.08	0.01	7.93	11,726	7,810
September, 1989	-0.10	0.08	0.02	4.61	3,082	2,441
March, 1990	-0.06	0.07	0.01	7.26	12,277	8,161
May, 1990	-0.02	0.08	0.01	8.84	1,426	9,927
September, 1990	-0.06	0.10	0.02	6.14	3,262	2,860
October, 1990	-0.01	0.10	0.01	13.01	14,207	9,896
March , 1991	0.00	0.13	0.01	14.09	14,232	9,601
May, 1991	0.03	0.10	0.01	13.43	14,711	10,346
September, 1991	-0.09	0.08	0.01	7.31	5,567	4,874
October, 1991	-0.04	0.10	0.01	12.32	14,609	10,678
March , 1992	0.00	0.12	0.01	15.71	15,229	10,248
May, 1992	0.04	0.10	0.01	13.65	15,388	10,422
August, 1992	-0.09	0.09	0.01	8.38	7,091	7,039
September, 1992	-0.12	0.06	0.01	4.31	4,133	3,803
October, 1992	-0.06	0.09	0.01	11.60	14,770	11,762
February, 1993	0.01	0.12	0.01	14.74	13,195	9,404
May, 1993	0.02	0.08	0.01	10.91	14,670	10,396
August, 1993	-0.07	0.09	0.01	8.04	6,928	7,014
September, 1993	-0.09	0.06	0.01	4.89	4,654	4,252
October, 1993	-0.04	0.13	0.01	16.71	15,150	11,451
Median	-0.04	0.10	0.01			
Mean	-0.04	0.09	0.01			
S.D.	0.05	0.02	0.00			

[a] Positive values indicate that undergraduate performance was superior to graduate performance. Mean differences shown here may differ slightly from results obtained from the TWE means in Table 5 because Table 5 results were conducted before data editing and because of rounding.

[b] All p-values were less than .001.

Table 9
Summary of All Group Comparisons - Region 1

Groups[a]	Number Valid[b]	Mean Difference[c] Percentiles 25th	50th	75th	SMD[c] Percentiles 25th	50th	75th	Number Extreme[d] <-.2	>+.2	SE(SMD) Percentiles 25th	50th	75th	Mantel Z Number Sig[e] -	+
G / UG	22	-0.11	-0.02	0.04	0.14	0.18	0.21	0	8	0.01	0.01	0.01	0	21
G: Hum/UG	15	-0.26	-0.22	-0.19	0.07	0.09	0.12	0	1	0.04	0.04	0.04	0	10
G: Soc. Sci./UG	18	-0.37	-0.29	-0.24	-0.03	0.04	0.06	0	0	0.02	0.02	0.03	1	11
G: Bio. Sci./UG	17	-0.33	-0.19	-0.10	0.00	0.11	0.16	1	2	0.03	0.03	0.04	3	12
G: Phys. Sci./UG	22	-0.49	-0.41	-0.24	-0.10	0.01	0.16	3	4	0.01	0.02	0.03	7	12
G: Bus/UG	20	-0.60	-0.48	-0.33	-0.09	-0.03	0.08	0	1	0.04	0.04	0.04	3	4
Hum/Soc. Sci.	15	0.00	0.05	0.10	-0.01	0.02	0.05	0	1	0.03	0.03	0.04	0	1
Hum/Bio. Sci.	15	-0.14	-0.03	0.10	-0.13	-0.09	0.06	1	1	0.03	0.03	0.04	7	3
Hum/Phys. Sci.	15	-0.03	0.06	0.19	-0.09	-0.04	0.08	1	1	0.03	0.03	0.04	3	4
Hum/Bus	15	0.17	0.24	0.30	0.06	0.11	0.15	0	1	0.04	0.04	0.05	0	7
Soc. Sci/Bio. Sci.	17	-0.13	-0.10	0.04	-0.12	-0.09	0.00	1	0	0.02	0.02	0.04	10	1
Soc. Sci./Phys. Sci.	18	-0.10	0.07	0.13	-0.10	-0.04	0.06	1	0	0.02	0.02	0.03	8	4
Soc. Sci./Bus	18	0.07	0.17	0.25	-0.02	0.07	0.09	0	1	0.03	0.03	0.04	0	7
Bio. Sci./Phys. Sci.	17	0.00	0.03	0.17	-0.01	0.02	0.05	0	0	0.02	0.02	0.04	1	3
Bio. Sci./Bus	17	0.11	0.24	0.34	-0.01	0.17	0.21	0	5	0.03	0.03	0.04	1	10
Phys. Sci./Bus	20	-0.04	0.14	0.24	-0.08	0.12	0.17	1	3	0.03	0.03	0.04	5	11

[a] G and UG denote graduate and undergraduate applicants, respectively.

[b] This gives the number of comparisons (out of 22) for which sample sizes were adequate for analysis.

[c] Positive values indicate superior performance by the group listed second.

[d] These columns give the number of SMDs (out of the Number Valid) that have magnitudes greater than .2.

[e] Values were considered statistically significant if they exceeded the critical value for a two-sided test at $\alpha = .01$. Separate counts are given for negative and positive Z values that were found significant.

Table 10

Summary of All Group Comparisons - Region 2

Groups[a]	Number Valid[b]	Mean Differences[c] Percentiles 25th	50th	75th	SMD[c] Percentiles 25th	50th	75th	Number Extreme[d] <-.2	>+.2	SE(SMD) Percentiles 25th	50th	75th	Mantel Z Number Sig[e] -	+
G / UG	22	-0.12	-0.03	-0.01	0.03	0.05	0.09	0	1	0.01	0.01	0.02	2	15
G: Hum/UG	9	-0.27	-0.25	-0.22	-0.08	-0.07	-0.03	0	0	0.05	0.05	0.05	1	0
G: Soc. Sci./UG	15	-0.36	-0.32	-0.27	-0.15	-0.11	-0.08	2	0	0.04	0.04	0.04	10	0
G: Bio. Sci./UG	9	-0.34	-0.19	0.01	-0.09	-0.04	0.07	1	1	0.03	0.05	0.05	3	1
G: Phys. Sci./UG	18	-0.16	0.00	0.06	0.10	0.13	0.18	0	1	0.03	0.03	0.04	3	13
G: Bus/UG	15	-0.28	-0.23	-0.15	-0.02	0.02	0.09	0	0	0.04	0.05	0.05	3	3
Hum/Soc. Sci.	9	-0.01	0.03	0.09	-0.06	0.02	0.09	0	0	0.06	0.06	0.06	0	0
Hum/Bio. Sci.	7	-0.11	0.02	0.05	-0.06	-0.02	0.07	1	0	0.06	0.07	0.07	1	1
Hum/Phys. Sci.	9	-0.28	-0.26	-0.14	-0.23	-0.16	-0.10	4	0	0.05	0.06	0.06	6	0
Hum/Business	9	-0.09	-0.07	0.02	-0.16	-0.14	-0.07	0	0	0.06	0.06	0.07	3	0
Soc. Sci/Bio. Sci.	9	-0.27	-0.09	-0.04	-0.15	-0.05	-0.04	1	0	0.05	0.06	0.06	2	0
Soc. Sci./Phys.Sci.	15	-0.34	-0.31	-0.25	-0.25	-0.23	-0.16	9	0	0.04	0.05	0.05	14	0
Soc. Sci./Business	15	-0.10	-0.04	-0.03	-0.19	-0.12	-0.07	4	0	0.05	0.05	0.06	6	0
Bio. Sci./Phys.Sci.	9	-0.20	-0.17	-0.05	-0.19	-0.11	-0.08	2	0	0.03	0.05	0.06	5	0
Bio. Sci./Business	9	0.00	0.02	0.20	-0.13	-0.08	-0.03	0	0	0.05	0.06	0.07	1	1
Phys Sci./Business	15	0.16	0.21	0.29	0.03	0.08	0.11	0	1	0.04	0.05	0.05	0	3

[a] G and UG denote graduate and undergraduate applicants, respectively.

[b] This gives the number of comparisons (out of 22) for which sample sizes were adequate for analysis.

[c] Positive values indicate superior performance by the group listed second.

[d] These columns give the number of SMDs (out of the Number Valid) that have magnitudes greater than .2.

[e] Values were considered statistically significant if they exceeded the critical value for a two-sided test at $\alpha = .01$. Separate counts are given for negative and positive Z values that were found significant.

Table 11
Summary of All Group Comparisons- Region 3

Groups[a]	Number Valid[b]	Mean Difference Percentiles 25th	Mean Difference Percentiles 50th	Mean Difference Percentiles 75th	SMD[c] Percentiles 25th	SMD[c] Percentiles 50th	SMD[c] Percentiles 75th	SMD[c] Number Extreme[d] <-.2	SMD[c] Number Extreme[d] >+.2	SE(SMD) Percentiles 25th	SE(SMD) Percentiles 50th	SE(SMD) Percentiles 75th	Mantel Z Number Sig[e] -	Mantel Z Number Sig[e] +
G / UG	22	-0.07	-0.04	0.00	0.08	0.10	0.10	0	0	0.01	0.01	0.01	0	22
G: Hum/UG	14	-0.27	-0.25	-0.20	0.04	0.06	0.09	0	0	0.04	0.05	0.05	0	0
G: Soc. Sci./UG	20	-0.35	-0.30	-0.27	0.00	0.03	0.05	0	0	0.03	0.03	0.04	1	0
G: Bio. Sci./UG	18	-0.10	-0.07	-0.05	0.11	0.16	0.19	0	3	0.04	0.04	0.05	0	18
G: Phys. Sci./UG	22	-0.18	-0.08	-0.05	0.11	0.15	0.18	0	4	0.03	0.03	0.04	0	21
G: Business/UG	19	-0.31	-0.26	-0.20	0.04	0.09	0.13	0	1	0.04	0.04	0.05	0	12
Hum/Soc. Sci.	14	0.01	0.03	0.09	-0.02	0.00	0.07	0	0	0.05	0.05	0.05	0	0
Hum/Bio. Sci.	14	-0.21	-0.16	-0.10	-0.15	-0.09	-0.08	0	0	0.05	0.06	0.06	4	0
Hum/Phys. Sci.	14	-0.21	-0.18	-0.09	-0.20	-0.14	-0.10	3	0	0.05	0.05	0.05	9	0
Hum/Business	14	-0.03	0.04	0.09	-0.12	-0.08	-0.01	0	0	0.05	0.05	0.06	1	0
Soc. Sci/Bio. Sci.	18	-0.26	-0.21	-0.16	-0.18	-0.13	-0.08	1	0	0.04	0.04	0.05	13	0
Soc. Sci./Phys. Sci.	20	-0.23	-0.20	-0.16	-0.18	-0.16	-0.12	2	0	0.03	0.03	0.04	18	0
Soc. Sci./Business	19	-0.07	-0.03	0.04	-0.10	-0.07	-0.05	1	0	0.04	0.04	0.05	5	0
Bio. Sci./Phys. Sci.	18	-0.03	0.01	-0.03	-0.06	-0.03	0.02	0	0	0.04	0.04	0.05	0	0
Bio. Sci./Business	18	0.13	0.18	0.24	0.01	0.06	0.11	0	0	0.04	0.05	0.05	0	2
Phys. Sci./Bus	19	0.12	0.18	0.29	0.05	0.07	0.11	0	1	0.04	0.04	0.05	0	6

[a] G and UG denote graduate and undergraduate applicants, respectively.

[b] This gives the number of comparisons (out of 22) for which sample sizes were adequate for analysis.

[c] Positive values indicate superior performance by the group listed second.

[d] These columns give the number of SMDs (out of the Number Valid) that have magnitudes greater than .2.

[e] Values were considered statistically significant if they exceeded the critical value for a two-sided test at $\alpha = .01$. Separate counts are given for negative and positive Z values that were found significant.

Cover Printed on Recycled Paper

57906-07593 • Y75M 5 • 275592 • Printed in U.S.A.